

REMARKS/ARGUMENTS

The amendment is in response to the Office Action dated January 14, 2005. Claims 1-23 are pending in the present application. Applicant thanks the Examiner for the interview on March 22, 2005. In response to the issues discussed in the interview, Applicant has amended claims 1, 2, 4-7, 9, 10, 12-15 and 17, and canceled claims 3 and 11. Accordingly, claims 1, 2, 4-10 and 12-23 remain pending in the present application.

Amended Claims

Independent claims 1, 9, and 17 were amended to recite temporarily storing the retrieved document in a storage device, determining whether relevant information is contained in the document, generating the document extract if the document contains relevant information and replacing the document in the storage device with the document extract. Support for this amendment is found in the claims and throughout the Specification for example at page 14, line 22 to page 15, line 15; page 16, line 20 to page 17, line 5; page 19, line 19 to page 20, line 1, and FIG. 3. Accordingly, no new matter has been presented.

In addition, claims 2, 4-7, 10, 12-15 were amended to provide consistent number references to the amended base claims. No new matter has been presented.

Claim Rejections

The Examiner rejected claims 1-6, 8-14, 16-20 and 22-23 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (U.S. Patent No. 6,631,369) in view of Nelson et al. (U.S. Patent No. 6,243,713). Claims 7, 15 and 21 were rejected under 35 U.S.C. §103(a) as

being unpatentable over Meyerzon in view of Nelson and further in view of Smadja. In rejecting the independent claims, the Examiner stated:

Regarding claims 1 and 9, Meyerzon discloses a method for retrieving information using a search engine comprising the steps of:

- (a) retrieving a document to be indexed (see col. 4, lines 43-54, Meyerzon);
- (b) generating a document extract corresponding to the document (see col. 4, lines 53-67); and
- (d) storing the plurality of tokens in a search index, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query (see col. 7, lines 44-65 and col. 8, lines 1-10, Meyerzon. The data type of information corresponding to the "token").

Meyerzon, however, does not explicitly disclose extracting a portion of the document that characterizes the document's subject content to form the document extract and decomposing the document extract into a plurality of tokens. Nelson, on the other hand, discloses the retrieval system for retrieval of multimedia information including the extracting a portion of the document and decomposing the document into a plurality of tokens (see abstract of Nelson; col. 5, line 52-col. 6, line 65, col. 7, lines 46-67 and col. 9, lines 60-65). It would have been obvious to one of ordinary skill in the art at the time of the invention to modify Meyerzon to include the claimed feature as taught by Nelson

Regarding claim 17, Meyerzon discloses a system for retrieving information, wherein the system includes a search engine comprising:

- means for retrieving a document from a documentary repository (see col. 4, lines 43-54 and element 200, Fig. 2 and corresponding text, Meyerzon);
- an information extractor coupled to the means for retrieving, wherein the information extractor generates a document extract corresponding to the document (see col. 4, lines 53-67, Meyerzon). Each document is retrieved from the web site process and the data is extracted from each of these retrieved documents. Therefore, there must be an extractor for the extracting process;
- a storage device (100, Fig. 2 and corresponding text, Meyerzon) coupled to the information extractor for storing the document extract;
- a search engine indexer (300, Fig. 2) coupled to the storage device; and
- a search index (400, Fig. 2) coupled to the search engine indexer for storing the plurality of tokens, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query (see col. 7, lines 44-65 and col. 8, lines 1-10; Fig. 2 and corresponding text, Meyerzon).

Meyerzon, however, does not explicitly disclose the steps of extracting a portion of the document that characterizes the document's subject content to form the document extract and decomposing the document extract into a plurality of tokens. Nelson, on the other hand, discloses the retrieval system for retrieval of

multimedia information including the decomposing the document into a plurality of tokens (see abstract of Nelson; col. 5, line 52-col. 6, line 65, col. 7, lines 46-67 and col. 9, lines 60-65). It would have been obvious to one of ordinary skill in the art at the time of the invention to modify Meyerzon to include the claimed feature as taught by Nelson

Applicant respectfully traverses.

The present invention relates to retrieving relevant data in large collections of documents. According to the present invention, a search index that reflects the characteristic portions of a document is created by utilizing an information extractor, which examines a document to determine if the document contains relevant information and if it does, generates a document extract. Conversely, if the document is irrelevant, e.g., the document is spam, the document is disregarded and a document extract is not generated. When an extract is generated, the document extract comprises only a portion of the document that is most characteristic of the document as a whole. Thus, the search index is based document extracts, and not on the document itself, and the document extracts are derived from documents containing relevant information.

Through aspects of the present invention, storage requirements are minimized because the document extract, and not the document, is stored in the search engine. The processing time to perform a document search is improved because the search index is smaller and does not contain references to inconsequential portions of a document. Finally, because the a document extract is generated only for relevant documents, the quality of the search result is improved.

The present invention, as recited in claim 1 provides:

1. A method for retrieving information using a search engine comprising the steps of:
 - (a) retrieving a document to be indexed and temporarily storing the document in a storage device;
 - (b) determining whether relevant information is contained in the document;
 - (c) if the document contains relevant information, generating a document extract corresponding to the document by extracting a portion of the

document that characterizes the document's subject content to form the document extract;

(d) replacing the document in the storage device with the document extract;

(e) decomposing the document extract into a plurality of tokens; and

(f) storing the plurality of tokens in a search index, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query.

Independent claims 9 and 17 are computer readable medium and system claims, respectively, having scopes similar to that of claim 1.

Independent claims 1, 9 and 17 are Allowable.

Applicant respectfully submits that none of the cited references, alone or in combination, teach or suggest the cooperation of elements recited in claims 1, 9 and 17. In particular, none of the references teaches or suggests determining whether relevant information is contained in the document and generating a document extract if the document contains relevant information, extracting a portion of the document that characterizes the document's subject content to form the document extract, and replacing the document in the storage device with the document extract, as recited in claims 1, 9 and 17. As stated above, a document extract is generated if the document contains relevant information. Moreover, the search index is based on the document extracts which characterize the subject content of the corresponding documents. Accordingly, the search index is based on the *semantic value* of the documents, as opposed to just the words or components of the document.

Meyerzon, is directed to minimizing the number of requests a web crawler makes to a document server to obtain the "increment" of the document set relative to the set of documents received during the previous crawl. Nelson is directed to indexing compound documents in a unified common index. In Nelson, a compound document, i.e., a document containing

multimedia components, is broken up into its constituent components (e.g., text, audio, images) and one or more tokens is created for each component. The components and their tokens are then stored in the unified common index (col. 2, lines 19-27).

While Meyerzon teaches “extracting the data from each of these retrieved documents and storing the data in an index” (column 4, lines 55-59), and Nelson teaches decomposing the compound document into its constituent multimedia components, indexing the components, and storing the indexed data in an index (column 5, lines 52-67), neither reference focuses on building an index based *relevant* documents and on the documents’ *subject content*. In particular, neither Meyerzon nor Nelson, singularly or in combination, teach or suggest “determining whether relevant information is contained in the document,” and “if the document contains relevant information, generating a document extract corresponding to the document by extracting *a portion of the document that characterizes the document’s subject content* to form the document extract” corresponding to the document and “replacing the document in the storage device with the document extract, ” as recited in claims 1, 9 and 17.

For the reasons presented above, Applicant respectfully submits that the cited references fail to teach or suggest the cooperation of elements recited in claims 1, 9 and 17 and that those claims are therefore allowable over the cited references. Claims 2, 4-8, 10, 12-16 and 18-23 depend on claims 1, 9 and 17, respectively, and the arguments above apply with full force to claims 2, 4-8, 10, 12-16 and 18-23. Accordingly, Applicant respectfully submits that claims 2, 4-8, 10, 12-16 and 18-23 are also allowable over the cited references.

Dependent Claims 5, 6, 13, and 14 are Allowable for Alternative Reasons

Applicant respectfully submits that dependent claims 5, 6, 13 and 14 are allowable over the cited references for reasons in addition to being dependent on allowable base claims. First,

neither reference teaches or suggests “extracting from the document a collection of sentences that are characteristic of the document’s subject content to form a document summary,” as recited in claims 5 and 13. In the Office Action, the Examiner states that Nelson teaches this feature at column 5, line 52 to column 6, line 65; column 7, lines 46-67 and column 9, lines 60-65. Those portions, however, discuss tokens and how that are generated. It mentions that “a text component (e.g., a paragraph of text) may be indexed by a number of tokens, each representing one or more words of the text component” (col. 6, lines 10-13), and that “a text token in most cases will represent an actual text string; e.g., the token ‘house’ will be used to index the word ‘house.’” (Col. 6, lines 17-19). Nothing in Nelson teaches or suggests “extracting from the document a collection of sentences *that are characteristic of the document’s subject content* to form a document summary,” as recited in claims 5 and 13.

Second, neither reference teaches or suggests “selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature,” as recited in claims 6 and 14. As discussed above, neither reference teaches or suggests generating the document extract. Therefore, it follows that neither reference can teach or suggest selecting any portion or part of the document extract. In the Office Action, the Examiner states that Nelson teaches this feature at column 6, lines 16-34, column 7, lines 46-67 and column 9, lines 60-65. Nevertheless, as discussed above, Applicant respectfully submits that the cited portions make no mention or suggestion of “selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature,” as recited in claims 6 and 14.

Conclusion

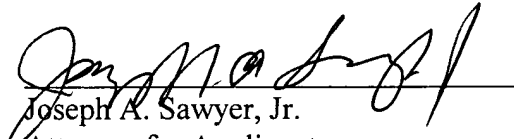
In view of the foregoing, Applicant submits that claims 1, 2, 4-10 and 12-23 are allowable over the cited references. Applicant respectfully requests reconsideration and allowance of the claims as now presented.

Applicant's attorney believes that this application is in condition for allowance. Should any unresolved issues remain, Examiner is invited to call Applicant's attorney at the telephone number indicated below.

Respectfully submitted,
SAWYER LAW GROUP LLP

April 8, 2005

Date



Joseph A. Sawyer, Jr.
Attorney for Applicant
Reg. No. 30,801
(650) 493-4540